Epidemiological Methods in Cancer Research

Molecular Cancer Epidemiology

Song Yao, PhD
Assistant Member

Department of Cancer Prevention & Control
Roswell Park Cancer Institute
Outline

• Cancer Epidemiology basics
• Molecular cancer epidemiology
• Integrative cancer epidemiology
“...the uniformity and predictability of many important biological phenomena take in the mass.”

- The principle of epidemiology by John Graunt 1662
Study Designs in Cancer Epidemiology

- Cross-Sectional Study
- Case-Control Study (Retrospective)
- Cohort Study (Prospective)
Case-Control Studies

- Type of analytic study in which participants are selected on the basis of whether they do or do not have the condition under investigation

- Compare cases and controls with respect to exposure of interest
Case-Control Studies and Cancer

- Suitable for rare diseases
  - participants selected with respect to disease status
  - feasible to obtain adequate sample size, even for rare cancer sites

- Potential for investigation of a wide variety of exposures
Case Definition and Selection

• Case definition should be as homogenous as possible
  – strict diagnostic criteria
    • incident cases vs. prevalent cases
    • *in situ* vs. invasive
    • disease subgroups (pre-vs. postmenopausal)

• Case selection should be tightly defined
  – all cases diagnosed at RPCI in 2006
  – all cases diagnosed in Buffalo in 2006
  – all cases diagnosed in WNY in 2006
Selection of Controls

- Controls should come from the population at risk for the disease or condition under investigation.
- Controls should be representative of the non-diseased populations.
- Controls should be like cases if they had developed the disease.
- Restrictions applied to case group should also be applied to control group.
Sources of Controls

• Hospital controls
  – recruit individuals with non-cancer conditions within same hospital
    • easy to identify, sufficient in numbers
    • ↓ recall bias (diseased like cases)
    • ↓ selection bias (same hospital)
    • ↓ non-response bias (motivated like cases)

• hospital controls are likely to differ from healthy population from which cases derived - greater prevalence of bad health habits
• many diseases inappropriate for control selection (shared risk factors)
Sources of Controls

• Community controls
  – when cases were selected to represent specified general population, controls can be drawn from that same population
  – random sample
    • random digit dialing
    • DMV, HCFA, voting lists, etc.
  – greatest level of comparability (same source population)
  – ↑ time and cost
  – ↑ recall bias (less motivated to recall exposure)
  – ↑ non-participation bias
Advantages

• quick and inexpensive

• well suited for diseases with long latency periods

• well suited for rare diseases

• allows for examination of multiple exposures
Disadvantages

- inefficient for rare exposures
- does not allow for computation of incidence rates
- particularly prone to bias
  - selection bias
  - recall bias
Bias

• Systematic error results in incorrect estimate of the association between exposure and disease
• Great potential for bias in epidemiologic studies - little control over participants
• Role of bias must be evaluated for interpretation of study results
• Bias, unlike confounding cannot be statistically controlled
• Major types: selection and information bias
Selection Bias

• Observed relation between exposure and disease is different among those who were selected for study, than for those who were eligible but were not selected

• Big problem in case-control studies: case and control selection based on different criteria, which are related to exposure status
Information (Observation) Bias

• Systematic error in the way exposure or disease data is obtained from study groups

• Case-control studies
  - recall bias: cases or controls may differ in recollection of exposure information
  - interviewer bias: differential probing for exposure history if disease status is known
Information (Observation) Bias (cont.)

- Misclassification
  - participants erroneously categorized with respect to either exposure or disease status
  - almost unavoidable
- nondifferential
  - extent of misclassification same for study groups
  - makes study groups more similar
  - risk estimate is biased toward the null
- differential
  - big problem
  - risk estimate can be overestimated, underestimated, or can change direction
Study Designs in Cancer Epidemiology

Exposure - Cross-Sectional Study - Disease

Exposure - Case-Control Study (Retrospective) - Disease

Exposure - Cohort Study (Prospective) - Disease
Cohort Studies

• Follow-Up Studies - Prospective Studies

• Individuals are defined on the presence or absence of exposure to a suspected risk factor

• Individuals are free from the disease under investigation at the time of exposure assessment

• Individuals are followed over a period of time to assess occurrence of disease
Prospective Cohort Studies

- Individual are defined on the basis of presence or absence of exposure to a suspected risk factor
- Exposure may or may not have occurred
- Outcome has not occurred
- Nurses’ Health Study, Physicians’ Health Study, Framingham Heart Study
Considerations in Cohort Enumeration

• Selection of exposed population
  – depends on exposure under investigation
  – common vs. rare

• Special exposure cohort
  – large proportion individuals exposed to agent

• Professional cohorts
  – good tracking system and information sources

• Population-based cohorts
  – generalizability
Selection of Comparison Group

• Comparison groups should be as similar as possible with respect to all other factors that may be related to the outcome except for the exposure of interest

• Internal comparison group
  – non-exposed cohort members

• External comparison group
  – general population
Follow-Up

- Cohort members followed from exposure (baseline) over time to determine whether or not they develop disease
- Incomplete follow-up on all cohort members major source of bias in cohort studies
- Length of follow-up determined by latency of outcome
  - cancer
  - birth defects
- Greater likelihood of bias with increased length of follow-up
Sources of Bias

• Less potential for bias compared to case-control studies
  – selection bias:
    • not likely - exposure assessed prior to disease onset
  – misclassification
    • potential for misclassification in exposure and disease status assessment
    • nondifferential misclassification
    • differential misclassification
Loss to Follow-Up

• Major source of bias in cohort studies
  – loss to follow-up related to exposure
  – loss to follow-up related to disease development

• Measures to counteract loss to follow-up should be implemented in design phase
  – engage participants in importance of research

• Assess differences between participants lost and those who remain in study

• Estimate risk under assumption that
  – all lost participants developed disease
  – no lost participants developed disease
Advantages

• Suitable for rare exposures

• Allows for investigation of multiple health effects associated with exposure

• Allows for investigation of temporal relationships

• Allows for computation of incidence rates
Disadvantages

• Inefficient for investigations of rare diseases

• Expensive and time consuming (prospective)

• Requires existing records (retrospective)

• Validity affected by loss to follow-up
Molecular Cancer Epidemiology
Definition

• An extension of traditional epidemiology that mandated incorporating biospecimens into epidemiologic study designs and enabled the merging of molecular, genetic and biochemical markers of exposure and/or early effect with questionnaire data

Malignant Progression of Human Cancer

Genetic Alteration
Environmental Exposure

Normal Tissue → Premalignant Lesions → Genetic Alteration → Primary Tumor → Genetic Alteration → Metastasis

- Normal Cell
- Premalignant Cell
- Malignant Cell without Metastatic Ability
- Malignant Cell with Metastatic Ability
Goals

• To understand mechanisms of carcinogenesis
• To refine cancer heterogeneity
• To identify high risk populations
• To predict outcomes or treatment response
• To identify biomarkers for early diagnosis or progression
• To discover novel targets for treatment or prevention
• To study gene-environment interactions
Required Resources

• Specimens – blood, marrow, buccal cells, normal/tumor tissue.....

• Advanced biotechnology equipment

• Team science: epidemiologists, pathologists, physicians, geneticists, biostatisticians and bioinformaticians, to integrate molecular biology into epidemiologic research and analyze its impact on clinical practice or public health

• Multi-disciplinary effort and collaboration between diverse fields
Important Concepts

• Mutation:
  – any change in the base pair sequence of genomic DNA
• Polymorphism:
  – a genetic trait where the least common allele is found in at least one percent of the population
• Genotype:
  – the genetic constitution of an individual
• Phenotype:
  – the appearance or other characteristics of an organism that result from the interaction of its genetic constitution with the environment
• Heterozygous:
  – different allelic genes at a locus in homologous chromosomes
• Homozygous:
  – identical allelic genes at a locus in homologous chromosomes
• Germline vs. Somatic
• Candidate targets vs. genome-wide
• Causal vs. Marker
• Driver vs. passenger
Genes

The functional and physical unit of heredity passed from parent to offspring. Genes are pieces of DNA, and most genes contain the information for making a specific protein.
<table>
<thead>
<tr>
<th>1st base</th>
<th>2nd base</th>
<th>3rd base</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>UUU (Phe/F) Phenylalanine</td>
</tr>
<tr>
<td>U</td>
<td>C</td>
<td>UCU (Ser/S) Serine</td>
</tr>
<tr>
<td>U</td>
<td>A</td>
<td>UAU (Tyr/Y) Tyrosine</td>
</tr>
<tr>
<td>U</td>
<td>G</td>
<td>UGU (Cys/C) Cysteine</td>
</tr>
<tr>
<td>C</td>
<td>U</td>
<td>UUC (Phe/F) Phenylalanine</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>UCC (Ser/S) Serine</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>UAC (Tyr/Y) Tyrosine</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>UGC (Cys/C) Cysteine</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
<td>UUA (Leu/L) Leucine</td>
</tr>
<tr>
<td>U</td>
<td>C</td>
<td>UCA (Ser/S) Serine</td>
</tr>
<tr>
<td>U</td>
<td>A</td>
<td>UAA Ochre (Stop)</td>
</tr>
<tr>
<td>U</td>
<td>G</td>
<td>UGA Opal (Stop)</td>
</tr>
<tr>
<td>C</td>
<td>U</td>
<td>UUG (Leu/L) Leucine</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>UCG (Ser/S) Serine</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>UAG Amber (Stop)</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>UGG (Trp/W) Tryptophan</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>CUU (Leu/L) Leucine</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>CCA (Pro/P) Proline</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>CAA (Gln/Q) Glutamine</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>CGU (Arg/R) Arginine</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>AUC (Leu/L) Leucine</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>CCG (Pro/P) Proline</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>CAG (Gln/Q) Glutamine</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>CGC (Arg/R) Arginine</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>AUA (Leu/L) Leucine</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>CCA (Pro/P) Proline</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>CAA (Gln/Q) Glutamine</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>CGA (Arg/R) Arginine</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>AUG (Met/M) Methionine</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>CCG (Pro/P) Proline</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>CAG (Gln/Q) Glutamine</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>CGG (Arg/R) Arginine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>AUU (Ile/I) Isoleucine</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>ACU (Thr/T) Threonine</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>AAU (Asn/N) Asparagine</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>AGU (Ser/S) Serine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>GUC (Ile/I) Isoleucine</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>ACC (Thr/T) Threonine</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>AAC (Asn/N) Asparagine</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>AGC (Ser/S) Serine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>GUA (Ile/I) Isoleucine</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>ACA (Thr/T) Threonine</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>AAA (Lys/K) Lysine</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>AGA (Arg/R) Arginine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>GUG (Val/V) Valine</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>GCA (Ala/A) Alanine</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>GAA (Glu/E) Glutamic acid</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>GGA (Gly/G) Glycine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>GUU (Val/V) Valine</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>GCC (Ala/A) Alanine</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>GAC (Asp/D) Aspartic acid</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>GGC (Gly/G) Glycine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>GUC (Val/V) Valine</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>GCA (Ala/A) Alanine</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>GAA (Glu/E) Glutamic acid</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>GGA (Gly/G) Glycine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>GUG (Val/V) Valine</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>GCG (Ala/A) Alanine</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>GAG (Glu/E) Glutamic acid</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>GGG (Gly/G) Glycine</td>
</tr>
</tbody>
</table>

**Categories:**
- Nonpolar
- Polar
- Basic
- Acidic
- Stop codon
Alternative Splicing

Translation

Protein A

Protein B

Protein C
Methods of Molecular Epidemiology

Human Genomes (normal or cancer)
- DNA sequence variations
- Gene expression
- Proteomics
- Metabolomics
- Epigenetics (methylation, acetylation, microRNAs, etc)

Other Genomes in Human Body
- Microbiome
Genetic Association Studies
Methods of Molecular Epidemiology

- DNA sequence variations
  - Single nucleotide polymorphisms
  - Structural variations
Single nucleotide polymorphism (SNP)
Human Genetic Variations

• 0.1% nucleotide differences = 1 per 1,000 base pairs = 3 million nucleotide differences

• Each person carries 250-300 loss-of-function variants in annotated genes

• Each person carries 50-100 variants previously implicated in inherited disorders

Diversity of human genetic background

Science 2007; 318:1842-1843.
SNP example

- **a** SNPs
- **b** Haplotypes
- **c** Tag SNPs
Linkage Disequilibrium

Mutation present on founder chromosomes

Population Expansion

Fragmentation of original chromosome by recombination

Critical Region

Linkage disequilibrium
Linkage Disequilibrium

rs10747524
rs11608702
rs12721364
rs7965281
rs10783215
rs7968585
rs2525046
Direct vs. indirect interrogation

a. Direct:

b. Indirect:

Establish A Reference Genome for SNPs

http://www.1000genomes.org/
http://www.1000genomes.org/
Genome-Wide Genotyping Array
## SNP Study Designs

<table>
<thead>
<tr>
<th>Candidate Gene</th>
<th>Genome wide scan</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Hypothesis-driven</td>
<td>• Exploratory</td>
</tr>
<tr>
<td>• Validate GWS</td>
<td>• No <em>a priori</em> knowledge</td>
</tr>
<tr>
<td>• False Negatives and Positives</td>
<td>• Hypothesis generating</td>
</tr>
<tr>
<td>• Require large (&gt;100s-1,000s) sample sizes to detect small effects</td>
<td>• Validate CGA</td>
</tr>
<tr>
<td>• Test gene-gene and gene-environment interactions</td>
<td>• Multiple Comparisons</td>
</tr>
<tr>
<td></td>
<td>• False Positives</td>
</tr>
<tr>
<td></td>
<td>• Limited ability to detect gene-gene or gene-environment interactions</td>
</tr>
<tr>
<td></td>
<td>• Require very large (&gt;1000s-100,000s) sample sizes</td>
</tr>
</tbody>
</table>
Workflow of genetic association analyses

Genotype data: format, storage

QC: sample/SNP call rate, gender/duplication/relatedness check, HWE, missingness, MAF

Categorization of genetic ancestry: reference, PCA, outlier

Imputation: reference, imputation probability

Single-locus analysis: genetic model (additive, dominant, recessive, genotypic, allelic)

Multi-locus analysis: haplotype, GxG interaction, aggregate genetic-risk score

Control for population admixture: genomic control, proportion of ancestry, PCs

Control for multiple comparison: Bonferroni (5x10^{-8}), permutation, FDR

Validation Validation Validation Validation
Example – Candidate Gene

Cdx2 (rs11568820)
Fok1 (rs10735810)
Bsm1 (rs1544410)
Apa1 (rs7975232)
Taq1 (rs731236)
Vitamin D Receptor

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*
**Figure 4 | Genome-wide scan for seven diseases.** For each of seven diseases $-\log_{10}$ of the trend test $P$ value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with $P$ values $<1 \times 10^{-5}$ highlighted in green. All panels are truncated at $-\log_{10}(P\text{ value}) = 15$, although some markers (for example, in the MHC in T1D and RA) exceed this significance threshold.
Published GWA Reports, 2005 – 6/2012

As of 01/25/13, the catalog includes 1,490 publications and 8,283 SNPs.

http://www.genome.gov/gwastudies/
Published Genome-Wide Associations through 07/2012
Published GWA at $p \leq 5 \times 10^{-8}$ for 18 trait categories
Challenges - Effect size vs. risk allele frequency

In published GWAS, median risk allele frequency = 0.36, the median OR = 1.33

Breast cancer

Challenges - Effect size vs. risk allele frequency

Performance of Common Genetic Variants in Breast-Cancer Risk Models

Sholom Wacholder, Ph.D., Patricia Hartge, Sc.D., Ross Prentice, Ph.D., Montserrat Garcia-Closas, M.D., Ph.D., Heather Spencer Feigelson, Ph.D., W. Ryan Diver, M.S.P.H., Michael J. Thun, M.D., David G. Cox, Ph.D., Susan E. Hankinson, Ph.D., Peter Kraft, Ph.D., Bernard Rosner, Ph.D., Christine D. Berg, M.D., Louise A. Brinton, Ph.D., Jolanta Lissowska, Ph.D., Mark E. Sherman, M.D., Rowan Chlebowski, M.D., Charles Kooperberg, Ph.D., Rebecca D. Jackson, M.D., Dennis W. Buckman, Ph.D., Peter Hui, B.S., Ruth Pfeiffer, Ph.D., Kevin B. Jacobs, B.S., Gilles D. Thomas, M.D., Robert N. Hoover, M.D., Sc.D., Mitchell H. Gail, M.D., Ph.D., Stephen J. Chanock, M.D., and David J. Hunter, M.B., B.S., Sc.D.

Copyright © 2010 Massachusetts Medical Society.
Performance of common genetic variants in breast cancer risk model

After adding 10 GWAS SNPs, the AUC of Gail model increased from 58.0% to 61.8%
Challenge - Synthetic Associations

Cirulli et al. Nat Rev Genet 2010;10:415
Sequencing Strategies

Figure 3 | Strategies for identifying disease-causing variants. Two discovery strategies using next-generation sequencing are likely to be of particular importance while sequencing costs remain high: sequencing in families with multiple affected individuals (a; shaded individuals are affected), and sequencing individuals at one or both ends of a trait distribution (b; the size of the individual represents the severity of the phenotype). In the case of family-based sequencing, it may often prove economical to first sequence the most distantly related co-affected individuals. Under either scenario, it is likely that follow-up genotyping in additional families or cohorts will be of particular importance to confirm the role of candidate variants. Red stars represent the causal variant. In a, stars of other colours represent variants that are shared by the sequenced individuals and do not segregate in the family.
## Challenge – Small Effect Rare Variants

**Table 1 | Potential frequencies of causal variants in complex traits**

<table>
<thead>
<tr>
<th>Variant class</th>
<th>Minor allele frequency</th>
<th>Implications for analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very common</td>
<td>Between 5 and 50%</td>
<td>Amenable to association analysis using current genome-wide association methods</td>
</tr>
<tr>
<td>Less common</td>
<td>Between 1 and 5%</td>
<td>Amenable to association analysis using variants catalogued in the 1000 Genomes Project</td>
</tr>
<tr>
<td>Rare (but not private)</td>
<td>Less than 1% but still polymorphic in one or more major human populations</td>
<td>Amenable to framework of extreme phenotype resequencing, as well as co-segregation in families</td>
</tr>
<tr>
<td>Private</td>
<td>Restricted to probands and immediate relatives</td>
<td>Difficult to analyse except through co-segregation in families. As linkage evidence will (by definition) be modest, discovery would be limited to the most recognizable of variants</td>
</tr>
</tbody>
</table>

*Cirulli et al. Nat Rev Genet 2010;10:415*
Gene-Environment Interactions
Smoking and breast cancer risk

- Inconsistent results of smoking and breast cancer risk in women
- NAT2 involved in the metabolism of aromatic amines

A meta-analysis 10 years later

<table>
<thead>
<tr>
<th>Smoking Status</th>
<th>Rapid NAT2</th>
<th>Slow NAT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never active smoker</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
</tr>
<tr>
<td>Pack-years &lt;20</td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
</tr>
<tr>
<td>Pack-years ≥20</td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
</tbody>
</table>

Sugar-Sweetened Beverages and Genetic Risk of Obesity

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Relative Risk</th>
<th>P Value for Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHS and HPFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;1 serving/mo</td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td>1–4 servings/mo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2–6 servings/wk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥1 servings/day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WGHS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;1 serving/mo</td>
<td></td>
<td>0.007</td>
</tr>
<tr>
<td>1–4 servings/mo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2–6 servings/wk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥1 servings/day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pooled</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;1 serving/mo</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1–4 servings/mo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2–6 servings/wk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥1 servings/day</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Relative Risk of the Development of Obesity per Increment of 10 Risk Alleles, According to Intake of Sugar-Sweetened Beverages.
Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals


Figure 5  Estimated absolute 10-year risk of colorectal cancer. 10-year absolute risk for (A) cancer-free men and (B) cancer-free women within the general population carrying >12, >13 and >14 risk alleles using 2006 Scottish population estimates (1 310 552 men, 1 441 245 women aged ≥35 years) using a Bayesian risk modelling approach. The rationale for assessing risk associated with carriage of various numbers of alleles is based on population frequency of that number of alleles and the associated risk (see figure 1). Ten years is taken as the predicted risk period because it is reasonable to expect colonoscopy to influence stage of colorectal cancer, mortality and/or incidence over that timescale. Cumulative probability is estimated from 1-exp(-cumulative rate) and the absolute risk in the next 10 years obtained by subtraction of the estimated cumulative risk up to the current age from the estimated cumulative risk for 10 years older than the current age. Risk is shown for men and women in each age group in the average risk population, subgroups with a positive family history and by genotype groups (note scale difference in plotting risks for men and women).
Methods of Molecular Epidemiology

- DNA sequence variations
  - Single nucleotide polymorphisms
  - Structural variations / copy number variations
Example – Copy Number Variants

A.

<table>
<thead>
<tr>
<th>LR</th>
<th>Background</th>
<th>Loss</th>
<th>Gain</th>
<th>CNN LOH</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>AAA</td>
<td>A</td>
<td>AAA</td>
<td>AAA</td>
</tr>
<tr>
<td>0.0</td>
<td>AA</td>
<td>A</td>
<td>AAA</td>
<td>AAA</td>
</tr>
<tr>
<td>0.5</td>
<td>AB</td>
<td>B</td>
<td>BBB</td>
<td>BB</td>
</tr>
<tr>
<td>1.0</td>
<td>BB</td>
<td>B</td>
<td>BBB</td>
<td>BB</td>
</tr>
</tbody>
</table>

B.

<table>
<thead>
<tr>
<th>LR</th>
<th>Background</th>
<th>Loss</th>
<th>Gain</th>
<th>CNN LOH</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>AAAA</td>
<td>AAA</td>
<td>AAAAA</td>
<td>AAAAA</td>
</tr>
<tr>
<td>0.0</td>
<td>AA</td>
<td>A</td>
<td>AAA</td>
<td>AAA</td>
</tr>
<tr>
<td>0.5</td>
<td>AABB</td>
<td>B</td>
<td>ABB</td>
<td>ABB</td>
</tr>
<tr>
<td>1.0</td>
<td>BBBB</td>
<td>B</td>
<td>BBB</td>
<td>BBB</td>
</tr>
</tbody>
</table>
Example – Copy Number Variants
Pharmacogenomics
Treatment Outcomes in Cancer Patients

- Despite treatments given, proportion of patients will not survive
  - Most emphasis, to date, on tumor tissue characteristics

- Variability noted in toxicities experienced with treatments

- Outcomes may be impacted by patient characteristics and behaviors

- Differential survival and toxicities could be related to inter-individual variability in drug metabolism and response to radiation
well every drug is going to have its adverse side effects!
Genetic markers for drug efficacy & toxicity

### Genetic markers for drug efficacy & toxicity

**Table:**

<table>
<thead>
<tr>
<th>Metabolism genotype</th>
<th>Receptor genotype</th>
<th>Response</th>
<th>Efficacy</th>
<th>Toxicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>65%</td>
<td>Low (5%)</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>32%</td>
<td>Low</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>9%</td>
<td>Low</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>79%</td>
<td>Moderate (15%)</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>40%</td>
<td>Moderate</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>10%</td>
<td>Moderate</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>80%</td>
<td>High (80%)</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>40%</td>
<td>High</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>10%</td>
<td>High</td>
</tr>
</tbody>
</table>

---

Pharmacogenetics/pharmacogenomics

- 20-95% variability in drug disposition and effects estimated to be attributed to genetics
- No selection pressure against variants important in drug efficacy or toxicity, thus common variants with larger effects expected
- Known single and potent causal factor
- Previous successes of pharmacogenetic markers
- Cancer pharmacogenetics: two genomes
PK/PD pathways of cyclophosphamide
Myeloperoxidase (MPO) genetic region

5' Promoter

-463 G/A (rs2333227)

SP1 binding sites

Coding exons

3' UTR

- Binding affinity
- Expression level
- ROS level

G allele

A allele
MPO variant rs2333227 and breast cancer treatment

HR=0.52 (0.29-0.92)

Treated group (CAF or CMF)

HR=0.41 (0.22-0.78)

Untreated group

HR=1.08 (0.50-2.31)


Cancer Genomics
The Landscape of Cancer Genome
Establishing A Reference Cancer Genome

- The Cancer Genome Atlas (TCGA)
- International Cancer Genome Consortium (ICGC)
ICGC Map – March 2012

47 projects launched

CANADA
- Pancreatic cancer (Ductal adenocarcinoma)
- Pediatric brain tumors (Medulloblastoma)
- Prostate cancer (Adenocarcinoma)

UNITED STATES
- Bladder cancer
- Blood cancer (Acute myeloid leukemia)
- Brain cancer (Glioblastoma multiforme/lower grade glioma)
- Breast cancer (Ductal & lobular)
- Cervical cancer (Squamous cell carcinoma)
- Colorectal cancer (Adenocarcinoma)
- Endometrial cancer (Uterine corpus endometrial carcinoma)
- Gastric cancer (Adenocarcinoma)
- Head and neck cancer (Squamous cell carcinoma/Thyroid carcinoma)
- Liver cancer (Hepatocellular carcinoma)
- Lung cancer (Adenocarcinoma/squamous cell carcinoma)
- Mesothelioma with AU/UK
- Prostate cancer (Adenocarcinoma)
- Renal cancer (Renal clear cell carcinoma/
- Renal papillary carcinoma)
- Skin cancer (Cutaneous melanoma)

EU/UNITED KINGDOM
- Breast cancer (ER positive, HER2 negative)
- Bone cancer (Osteosarcoma/chondrosarcoma/rare subtypes)
- Breast cancer (Triple negative/fibular/other)
- Chronic Myeloid Disorders (Myelodysplastic syndromes, myeloproliferative neoplasms and other chronic myeloid malignancies)
- Esophageal cancer
- Mesothelioma with AU/UK
- Prostate cancer

EU/FRANCE
- Renal cancer (Renal cell carcinoma) (Focus on but not limited to clear cell subtype)
- Breast cancer (Subtype defined by an amplification of the HER2 gene)
- Liver cancer (Hepatocellular carcinoma) (Secondary to alcohol and adiposity)
- Prostate cancer (Adenocarcinoma)

FRANCE
- Breast cancer (Subtype defined by an amplification of the HER2 gene)
- Liver cancer (Hepatocellular carcinoma) (Secondary to alcohol and adiposity)
- Prostate cancer (Adenocarcinoma)

GERMANY
- Malignant lymphoma (Germinal center B-cell derived lymphomas)
- Pediatric brain tumors (Medulloblastoma and Pediatric pilocytic astrocytoma)
- Prostate cancer (Early onset)

SAUDI ARABIA
- Thyroid cancer (Papillary carcinoma)

CHINA
- Gastric cancer (Intestinal- and diffuse-type)

JAPAN
- Liver cancer (Hepatocellular carcinoma) (Virus-associated)

SOUTH KOREA
- Breast cancer

MEXICO
- Blood cancer (Diffuse large B-cell lymphoma)
- Breast cancer (Ductal carcinoma)
- Cervical cancer
- Head and Neck Cancer (Squamous cell carcinoma of oral cavity/hypopharynx/larynx)
- Pediatric solid tumors

ITALY
- Rare pancreatic tumors (Enteropancreatic endocrine tumors and rare pancreatic exocrine tumors)

INDIA
- Oral cancer (Gingivobuccal)

AUSTRALIA
- Mesothelioma with AU/UK
- Ovarian cancer (Serous cystadenocarcinoma)
- Pancreatic cancer (Ductal adenocarcinoma)
- Prostate cancer
The Breast Cancer Genome Landscape

ARTICLE
Whole-genome analysis informs breast cancer response to aromatase inhibition

LETTER
The landscape of cancer genes and mutational processes in breast cancer

LETTER
The clonal and mutational evolution spectrum of primary triple-negative breast cancers

LETTER
Sequence analysis of mutations and transcriptomic processes across breast cancer subtypes

ARTICLE
Comprehensive genomic characterization defines human glioblastoma genes and core pathways

ARTICLE
The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups
From Cancer Genomics to Personalized Medicine
Methods of Molecular Epidemiology

- Epigenetics (methylation, acetylation, microRNAs, etc)
Epigenetics
Example - Epigenetics

http://www.med.ufl.edu/biochem/keithr/research.html
DNA Methylation Variations

Epigenome-Wide Association Study

<table>
<thead>
<tr>
<th>Type</th>
<th>Key advantage</th>
<th>Key disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case versus control (singleton)</td>
<td>Many cohorts exist</td>
<td>Cannot easily control for environmental and genetic confounders</td>
</tr>
<tr>
<td>Families</td>
<td>Can study potential inheritance</td>
<td>Few large cohorts of this type exist</td>
</tr>
<tr>
<td>Disease-discordant monozygotic twins</td>
<td>Can control for genetics</td>
<td>Few large cohorts of this type exist</td>
</tr>
<tr>
<td>Prospectively sampled, longitudinal</td>
<td>Can establish causality</td>
<td>Slow and difficult to establish</td>
</tr>
</tbody>
</table>

Methods of Molecular Epidemiology

- Gene expression
Breast Cancer Intrinsic Subtypes

JAMA. 2006;295:2492-2502
Intrinsic molecular subtypes

Carey 2006; JAMA 295:2492-2502
Methods of Molecular Epidemiology

- Proteomics
- Metabolomics
Detection of Bladder Cancer in Human Urine by Metabolomic Profiling Using High Performance Liquid Chromatography/Mass Spectrometry


From the Laboratory of Proteomics and Analytical Technologies (HJI, TW, TDV) and Advanced Biomedical Computing Center (BL), SAIC Frederick, Inc., NCI-Frederick, Frederick, Maryland, and Department of Urology, Bnai-Zion Medical Center (ON, EJI, AK, MM), Haifa, Israel
Example - Metabolomics

Normal Controls

Bladder Cancer

J Urology Vol 179, 2422-2426, June 2008
Example - Metabolomics

PCA of HPLC-MS data on 48 control and 41 bladder cancer urine samples

J Urology Vol 179, 2422-2426, June 2008
Methods of Molecular Epidemiology

Human Genomes (normal or cancer)
- DNA sequence variations
- Gene expression
- Epigenetics (methylation, acetylation, microRNAs, etc)

Other Genomes in Human Body
- Microbiome

Proteomics
Metabolomics
Human Microbiome Project

http://huttenhower.sph.harvard.edu/metaphlan
Diversity of Human Microbiome
Example – Oral Microbiome and Pancreatic Cancer Risk
Integrative Epidemiology

• A cohesive approach to combine the rigor of epidemiologic study design with rapid advances in analytic systems and biostatistical and bioinformatic tools, using the same populations, biospecimens, and data elements as in case-control or cohort studies of risk to extend to studies of outcome and response to therapy, as well as cancer risk-taking behaviors.

Spitz et al. Cancer Discover 2012
Contact Info

song.yao@roswellpark.org

x4968